Subspace Polynomials and List Decoding of Reed-Solomon Codes

by

Swastik Kopparty

Submitted to the Department of Electrical Engineering and Computer Science
in partial fulfillment of the requirements for the degree of

Master of Science in Computer Science and Engineering

at the

MASSACHUSETTS INSTITUTE OF TECHNOLOGY
[February 2007]
December 2006

Author .....
Department of Electrical Engineering and Computer Science
December 13, 2006

Certified by .....
Madhu Sudan
Professor of Computer Science
Thesis Supervisor

Accepted by .....
Arthur C. Smith
Chairman, Department Committee on Graduate Students

Subspace Polynomials and List Decoding of Reed-Solomon Codes

by

Swastik Kopparty

Submitted to the Department of Electrical Engineering and Computer Science
on December 13, 2006, in partial fulfillment of the
requirements for the degree of
Master of Science in Computer Science and Engineering

Abstract

We show combinatorial limitations on efficient list decoding of Reed-Solomon codes beyond the Johnson and Guruswami-Sudan bounds [Joh62, Joh63, GS99].

In particular, we show that for any $\delta \in (0,1)$, there exist arbitrarily large fields $\mathbb{F}_N, |\mathbb{F}_N| = N$, such that for $K = N^\delta$:

- **Existence:** there exists a received word $w_N : \mathbb{F}_N \to \mathbb{F}_N$ that agrees with a super-polynomial number of distinct degree K polynomials on $\approx N^{\sqrt{\delta}}$ points each;
- **Explicit:** there exists a polynomial time constructible received word $w'_N : \mathbb{F}_N \to \mathbb{F}_N$ that agrees with a super-polynomial number of distinct degree K polynomials, on $\approx 2^{\sqrt{\log N}} K$ points each.

In both cases, our results improve upon the previous state of the art, which was $\approx N^\delta/\delta$ for the existence case [JH01], and $\approx 2N^\delta$ for the explicit one [GR05b].

Furthermore, for δ close to 1 our bound approaches the Guruswami-Sudan bound (which is $\sqrt{NK}$) and rules out the possibility of extending their efficient RS list decoding algorithm to any significantly larger decoding radius.

Our proof method is surprisingly simple. We work with polynomials that vanish on subspaces of an extension field viewed as a vector space over the base field. These *subspace polynomials* are a subclass of *linearized polynomials* that were studied by Ore [Ore33, Ore34] in the 1930s and by coding theorists. For us their main attraction is their sparsity and abundance of roots.

We also complement our negative results by giving a list decoding algorithm for linearized polynomials beyond the Johnson-Guruswami-Sudan bounds.

Thesis Supervisor: Madhu Sudan
Title: Professor of Computer Science

Acknowledgments

I am grateful to Madhu, my advisor, for advice, support, perspective and for many explanations and discussions on research matters. The results in this thesis are joint with Eli Ben-Sasson and Jaikumar Radhakrishnan. Many thanks to Eli and Jaikumar for the constant encouragement.

Swastik Kopparty

Fall 2006

Contents

Chapter 1

Introduction

1.1 Error-correcting codes and Combinatorial list decoding

An *error-correcting code* is a combinatorial object designed to enable reliable transmission of information on a noisy channel. A fundamental task associated with the use of error-correcting codes is *decoding,* that is, recovering the original codeword w, from the corrupted received word w'. In *list decoding*, introduced independently by Elias [Eli57] and Wozencraft [Woz58], we consider a relaxation of the above task. We no longer demand that the codeword w be recovered uniquely. Instead, we ask for a *list* of possible codewords, and we call the recovery successful if w appears in this list. The crucial point is that list decoding (with relatively small lists) is possible even when the received word's agreement with the original word is so small that any hope of recovering the original word uniquely is unrealistic.

Clearly, the size of the list is a crucial parameter of the problem (to wit, the set of all codewords is a trivial solution for unbounded list size). In order to quantify the number of errors one can tolerate for list decoding and appreciate the difference with unique decoding, let us consider a code with block length N and distance d. Suppose the number of errors in the transmission is guaranteed to be at most e. Then, in the worst case, we can uniquely recover the original message only if e is at most $\frac{d-1}{2}$.

Thus, even if d is very close to N, we require that the two words agree on at least $N - \frac{d-1}{2} \geq \frac{N}{2}$ places. With list decoding we can, in principle, tolerate substantially more errors.

Theorem 1.1.1 (Johnson bound [Joh62, Joh63, GS01]). *Let $\mathcal{C}$ be a code with block length N and distance d. Then, the number of codewords that agree with any word w' on more than $\sqrt{N(N-d)}$ places is $O(N^2)$.*

For example, if the code has distance $0.99N$, then there are only a polynomial number of words that agree with the given word on more than $\frac{N}{10}$ places, that is, we have polynomial size lists as long as the error is less than $0.9N$. In fact, a more precise form of the Johnson bound [GS01] implies that if the error is less than $0.8N$, then there are only a constant number of such codewords. In contrast, unique decoding is feasible only when the error is guaranteed to be less than $0.495N$.

Given a specific family of codes (in this paper, we shall consider Reed-Solomon codes), we recall two fundamental problems associated with list decoding:

- **Combinatorial list decoding:** Given code $\mathcal{C}$ of block-length N and agreement parameter $A \leq N$, estimate the maximal number of distinct codewords that agree with w' on at least A entries, where the maximum is taken over all (possibly corrupted) received words w'.

- **Algorithmic list decoding:** Devise an efficient algorithm for $\mathcal{C}$ that on input w' and agreement parameter A, lists all words that agree with w' on at least A entries.

In order to solve the algorithmic list decoding problem (for a specific code $\mathcal{C}$ and agreement parameter A) in worst-case time T, at the very least we should have a combinatorial guarantee that there are no more than T codewords in the list. In other words, lower bounds for the combinatorial problem imply lower bounds for the algorithmic one.

Our main result (Theorem 2.0.1) is an improved lower bound for the combinatorial list decoding problem for Reed-Solomon codes (described next). This result implies

super-polynomial lower bounds on the running time of any list decoding algorithm for Reed-Solomon codes for a sufficiently small agreement parameter.

1.2 Reed-Solomon codes

Reed-Solomon codes, named after their inventors [RS60], are an extensively studied family of error correcting codes, heavily used in theoretical and practical settings. The codewords of the Reed-Solomon (RS) code over a field $\mathbb{F}$ are the evaluations of low degree polynomials at N distinct field elements. Bounding the degree of polynomials by K yields a $[N, K+1, N-K]_{\mathbb{F}}$ code, i.e. a linear code over alphabet $\mathbb{F}$ with block-length N, dimension $K+1$ and distance $N-K$. The optimality of the RS codes (in terms of dimension to distance tradeoff) and their many algebraic properties contributed to the usefulness and ubiquity of RS codes in coding theory and theoretical computer science. In what follows we will restrict our attention to RS codes evaluated over the whole field of size N (i.e. $\mathbb{F} = \mathbb{F}_N$). Thus, in the space of all possible received words $\{w : \mathbb{F}_N \to \mathbb{F}_N\}$, the RS code we consider is the subset

$$\mathsf{RS}[N, K] := \{\langle P(\alpha) : \alpha \in \mathbb{F}_N \rangle : P \text{ is a polynomial with } \deg(P) \leq K\}.$$

In this context, the pair of fundamental problems mentioned above boil down to:

- **Combinatorial RS list decoding:** Given field $\mathbb{F}_N$, degree K and agreement parameter $A \leq N$, bound the maximum, over all $w : \mathbb{F}_N \to \mathbb{F}_N$, of the number of polynomials $P \in \mathsf{RS}[N, K]$ with $\mathsf{agree}(w, P) \geq A$, where $\mathsf{agree}(f, g) := |\{x \in \mathbb{F}_N : f(x) = g(x)\}|$

- **Algorithmic RS list decoding:** Given $\mathbb{F}_N, K, A$ and received word w as above, efficiently produce the list $\{P \in \mathsf{RS}[N, K] : \mathsf{agree}(P, w) \geq A\}$.

Following a breakthrough by Sudan [Sud97], Guruswami and Sudan [GS99] presented a polynomial time algorithm for solving the algorithmic RS list decoding problem for agreement parameter $A > \sqrt{KN}$. The Guruswami-Sudan algorithm has found spectacular applications not only to coding theory, but also to complexity

theory and derandomization (see the surveys [GS02, Gur06] and pointers therein). Interestingly, the minimal agreement needed for the Guruswami-Sudan list decoder to succeed coincides with the Johnson bound stated above. Following the list decoding algorithm of Guruswami-Sudan for Reed-Solomon codes, list decoding algorithms for several other codes have been discovered. Recent breakthroughs of Parvaresh & Vardy [PV05] and Gurswami & Rudra [GR05b] have lead to the design of codes (related to Reed-Solomon codes) where efficient list decoding is possible well beyond the Johnson bound. For Reed-Solomon codes itself, however, the Guruswami-Sudan decoder still provides the best bounds. Since the recent revolution in the understanding of list decoding started with Reed-Solomon codes, and subsequent developments have rested on it, and given the ubiquity of RS codes in theory and practice, one would like to obtain a precise understanding of their list decoding properties. In particular, we would like to know how big we can allow the radius of the ball to be (or how small the agreement can be) and still guarantee polynomially many codewords in it. Additionally, lower bounds for list decoding, especially when the received "bad" word w (that agrees with many RS-codewords) can be produced in polynomial time in N, have applications for hardness of approximating the minimum distance of a linear code [DMS99] and constructing error-correcting codes with improved parameters [Xin03].

Prior to this work, there have been several attempts to show combinatorial lower bounds on the agreement parameter for efficient list decoding. However, despite progress on related combinatorial problems (such as RS list recovery [GR05b] and RS list hitting [CW04], see Section 2.4), the state of the art regarding the combinatorial RS list decoding problem was rather weak. Apart from the trivial observation that list decoding to K agreements could produce superpolynomial size lists, the only bound followed from a relatively straightforward combinatorial argument due to Justesen & Høholdt [JH01] showing (only) for low rates $K = N^{\delta}$ $(0 < \delta < 1)$ that super-polynomially many words agree with some received word on $\approx N^{\delta}/\delta$ entries. In this paper, we significantly improve on this state of affairs and show that super-polynomial list size can be obtained for much higher agreement, that at certain parameters even approaches the Johnson bound. Before describing our results, we

present a reformulation of the original argument and motivate the use of subspace polynomials in improving it.

1.3 Overview of our technique: viva subspace polynomials

Let us start by sketching an argument that shows that for $K = N^\delta$ $(0 < \delta < 1)$, at least $\frac{1}{\delta}N^\delta$ agreements are required for efficient list decoding. Consider all monic univariate polynomials in X of degree T that have exactly T roots in $\mathbb{F}_N$. There are exactly $\binom{N}{T}$ polynomials. Of these, at least $\binom{N}{T}N^{T-K-1}$ have the same coefficients for monomials of degree greater than K. Construct the received word w by evaluating one of these polynomials P^* at all points in $\mathbb{F}_N$. Now if P' is another such polynomial, then $P^* - P'$, a polynomial of degree at most K, agrees with w on T places (because $P^* - (P^* - P') = P'$ has T roots). Thus, in order to have polynomial size lists, $\binom{N}{T}N^{-(T-K-1)}$ must be bounded by a polynomial. This implies that $T > \frac{1}{\delta}N^\delta$.

Let us abstract out what was important in the above argument: a large set of polynomials with many roots that agree on their top coefficients. The key idea in our construction is to start with sparse polynomials. The reason this helps is that we lose fewer polynomials when we try to get their top coefficients to agree. So are there sparse polynomials of small degree that have many roots? This is where subspace polynomials come in. Let us briefly describe what they are. Fix a base field $\mathbb{F}_q$ and view its extentsion $\mathbb{K} = F_{q^m}$ as a vector space over $\mathbb{F}_q$. Let L be a T element subspace of this vector space. The subspace polynomial associated with L is $P_L(X) = \prod_{\alpha \in L}(X - \alpha)$. It turns out that such polynomials are sparse: only monomials of the form X^{q^i} have non-zero coefficients. Thus, a fraction $N^{-\log_q(T/K)}$ of them have the same top coefficients. This is much less than the factor $N^{-(T-K-1)}$ that we had in the previous argument, however since we require that the roots form a subspace, we do not start with $\binom{N}{T}$ polynomials but only as many as there are T element subspaces of $\mathbb{K}$. However, as we will see, the tradeoff between these numbers

is favorable, and we do obtain a better bound overall.

Subspace polynomials are a special class of *linearized* polynomials. These are polynomials that represent $\mathbb{F}_q$-linear functions from $\mathbb{K}$ to $\mathbb{K}$, and they are also sparse in exactly the same way as subspace polynomials. These polynomials were first studied by Ore [Ore33, Ore34] and have important uses in the study of finite fields and applications (see Berlekamp [Ber68] and Lidl and Niederreiter [LN97, Chapter 3, Section 4]). Subspace polynomials recently made their first appearance (to the best of our knowledge) in the context of computational complexity in the construction of short PCPs and their related locally testable codes due to Ben-Sasson, Goldreich, Harsha, Sudan and Vadhan [BSGH+04]. They also play (several different) pivotal roles in the short PCPs of [BSS05] and in sub-linear proof verification of [BSGH+05]. The properties of subspace polynomials we exploit here are similar to those exploited by the aforementioned works (but the context is completely different)— their sparsity, their many roots and the fact that they represent linear transformations over a vector space.

Organization In the Chapter 2 we state the main result and discuss its corollaries. Chapter 3 contains a proof of the main theorem. In section 3.2 we give a list decoding algorithm for linearized polynomials. We conclude with some open problems.

Chapter 2

The main result and its corollaries

We construct a large family of words from $\mathsf{RS}[N, K]$ that reside in a ball of small radius. We will work in the field $\mathbb{F}_{q^m}$, that is $N = q^m$. K will be of the form q^u for integer u $(0 \leq u \leq m)$. We start by stating our result in its general form. Later, we specialize this general result and compare it with the previous state of the art.

Theorem 2.0.1 (Main Theorem). *Let q be a prime power and m a positive integer. Let u and v be integers such that $0 \leq u \leq v \leq m$. Then, there is a family $\mathcal{P} \subseteq \mathbb{F}_{q^m}[X]$ of polynomials of degree q^u and a word $w : \mathbb{F}_{q^m} \to \mathbb{F}_{q^m}$ such that*

1. $|\mathcal{P}| \geq q^{(u+1)m-v^2}$;
2. *for all $P \in \mathcal{P}$, $\mathsf{agree}(w, P) \geq q^v$;*
3. *there exist scalars $\alpha_{u+1}, \alpha_{u+2}, \ldots, \alpha_{v-1} \in \mathbb{F}_{q^m}$ such that for all $x \in \mathbb{F}_{q^m}$,*

$$w(x) = x^{q^v} + \sum_{i=u+1}^{v-1} \alpha_i x^{q^i}.$$

By choosing q, m, u and v appropriately we can obtain several interesting corollaries of this construction; we describe these corollaries below.

2.1 Low rate

As stated above, Justesen and Høholdt [JH01] obtained their lower bound for the case, $K = N^\delta$ $(0 < \delta < 1)$. Using our main theorem we can show super-polynomial

number of codewords even with substantially larger agreement.

Corollary 2.1.1 (Low rate). *Let $\delta, \rho \in (0,1)$ be rational numbers such that $\delta < \rho$. Then, for infinitely many N, there is a word $w : \mathbb{F}_N \to \mathbb{F}_N$, such that*

$$\left|\{P \in \mathsf{RS}[N, N^\delta] : \mathsf{agree}(w, P) \geq N^\rho\}\right| \;\geq\; N^{(\delta-\rho^2)\log_2 N}.$$

Proof. Let $q = 2$ and let m be large enough such that $u = \delta m$ and $v = \rho m$ are integers. Our claim follows immediately from Theorem 2.0.1. □

Comparison to earlier results: (See Figure 2-1) The previous state of the art for $K = N^\delta$, was that there are super-polynomially many codewords if the agreement required is $\frac{N^\delta}{\delta}$ [JH01] (see [CW04] for a related argument). For instance, for $\delta = \frac{1}{2}$ this quantity is $2\sqrt{N}$. Corollary 2.1.1 improves the super-polynomial lower bound on list size for agreement as large as $N^{\sqrt{\delta}-\epsilon}$ for any $\epsilon > 0$. (Note, however, that the the previous arguments gave lists of size $2^{N^{\Omega(1)}}$, whereas our result only promises lists of size $2^{\Omega(\log^2 N)}$.) As for upper bounds, the Johnson bound says that with agreement greater than $N^{(1+\delta)/2}$ we obtain only polynomially large lists. Once again, for $\delta = \frac{1}{2}$, the Johnson bound limits the number of codewords at agreement parameter $N^{0.75}$, whereas Corollary 2.1.1 states that reducing the agreement to $N^{0.7}$ results in lists of size $N^{\Omega(\log N)}$.

Notice that when δ is close to 1, say $\delta = 1 - \gamma$ (for some small constant γ), then the Johnson bound ensures that there are only $O(N^2)$ words if we require an agreement more than $N^{1-\frac{\gamma}{2}}$, whereas Corollary 2.1.1 shows that there are super-polynomially many words if we relax the requirement to only $N^{1-\frac{\gamma}{2}-\frac{\gamma^2}{4}}$, (because $(1-\frac{\gamma}{2}-\frac{\gamma^2}{4})^2 < 1-\gamma$, for $\gamma < \frac{1}{2}$).

2.2 Constant rate

Corollary 2.2.1 (Constant rate). *Let r' and r be positive integers such that $r' \leq r \leq 2r'$. Let $R = 2^{-r}$ and $R' = 2^{-r'}$. Then, for infinitely many N, there is a word*

$w : \mathbb{F}_N \to \mathbb{F}_N$, *such that*

$$\left|\{P \in \mathsf{RS}[N, RN] : \mathsf{agree}(w, P) \geq NR'\}\right| \;\geq\; N^{2r'-r}.$$

Proof. This time we apply Theorem 2.0.1 with $q = 2$, m large, $u = m - r$ and $v = m - r'$. □

Comparison to earlier results: No non-trivial bounds were known for the case of high rate. In particular, the counting arguments used for the case of low rate do not say anything significant. [As before, there are at least $\binom{N}{T}/N^{T-K}$ codewords from $\mathsf{RS}[N, K]$ that all agree with some word on T places. However, if we choose $K = RN$ for some constant R, then in order to make $\binom{N}{T}/N^{T-K}$ superpolynomial, we need to restrict T to be $K + O(\frac{N}{\log N}) = RN(1 + o(1))$].

As for the upper bounds for rate R, the Johnson bound shows that the list size is $O(N^2)$ when the agreement is $\geq \sqrt{R}N$. In contrast, Corollary 2.2.1 states that if the radius is decreased to say $R^{\frac{1}{2}+\epsilon}N$ (this corresponds to choosing $r' = \left(\frac{1}{2} + \epsilon\right) r$) then the list size cannot be bounded by a fixed polynomial independent of R: it must be at least $N^{\epsilon \log(\frac{1}{R})}$.

For the sake of completeness and due to the importance of algorithmic RS list decoding, we explicitly state the following limitation to it. Recall that the Guruswami-Sudan algorithm finds a list of all codewords with agreement at least $(NK)^{\frac{1}{2}}$ with the received word. On the other hand it was known that list-decoding to agreement K could encounter superpolynomial sized lists. The following statement says that one cannot get a significantly better list decoding algorithm that works in as general as setting as the Guruswami-Sudan algorithm.

Corollary 2.2.2 (Limits of RS algorithmic list decoding). *For all $\epsilon > 0$, there is no polynomial time algorithm that, for* any *N, K and received word $w : \mathbb{F}_N \to \mathbb{F}_N$, produces a list of all $P \in RS[N, K]$ with* $\mathsf{agree}(w, P) > K^{\frac{1}{2}+\epsilon} N^{\frac{1}{2}-\epsilon}$.

Remark: Note however that the corollary produces counterexamples for very low rates only. It may still be the case that for high rates one can list-decode to large radii.

2.3 Explicit constructions

The lower bounds stated above imply the existence of a word w in large agreement with many RS-codewords. However, we do not claim w can be constructed in polynomial time in N (inspection of Part 3 of Theorem 2.0.1 reveals it can be computed in quasi-polynomial time). Restricting our attention to efficient and explicit constructions, we obtain the following.

Corollary 2.3.1 (Explicit construction). *Let $\delta \in (0,1)$. Let t and N be such that δt is an integer and $q = N^{\frac{1}{t}}$ is a prime power. Then, over the field $\mathbb{F}_N$, we have*

$$\left|\{P \in \mathsf{RS}[N, N^{\delta}] : \mathsf{agree}(w, P) \geq qN^{\delta}\}\right| \;\geq\; q^{(\delta t+1)(t-\delta t-1)},$$

where $w(x) = x^{q^{\delta t+1}}$ for $x \in \mathbb{F}_N$.

Proof. We apply Theorem 2.0.1 with $q = N^{\frac{1}{t}}$, $u = \delta t$ and $v = \delta t + 1$. Note that since $v = u + 1$, the polynomial $P^*(X) = X^{q^v}$ can be computed explicitly, as claimed. □

Comparison to earlier results: The previous state of the art, due to the recent result of Guruswami and Rudra [GR05b], constructed an explicit word w with $2^{N^{\Omega(1)}}$ many codewords with agreement $(2-\epsilon)N^{\delta}$ with w. For appropriate choice of parameters, Corollary 2.3.1 can (for example) give an explicit w with $N^{\Omega(\log N)}$ codewords having agreement $2^{\sqrt{\log N}}N^{\delta}$. Thus, the agreement in our result is significantly larger than that of [GR05b], yet the number of codewords is significantly smaller. We stress that although we do get a super-polynomial number of codewords, for application to [DMS99] one needs $2^{N^{\Omega(1)}}$ codewords with much larger agreement.

2.4 Other Related Work

There are a few results about problems related to the combinatorial RS list decoding problem, that we now describe. Guruswami and Rudra [GR05b] prove that the Guruswami-Sudan list decoding algorithm is optimal for a more general problem called Polynomial-Reconstruction and a related problem called List-Recovering of

Reed-Solomon codes. Cheng and Wan [CW04] prove that if there is a polynomial time algorithm for list decoding of Reed-Solomon codes up to some radius, then certain cryptographic assumptions would be false. The results in this paper make this statement vacuous over any fields where our constructions apply. Kiayias and Yung [KY02] and others have based cryptographic schemes on the hardness of list decoding Reed-Solomon codes. Our work proves results in the direction of validating their assumptions.

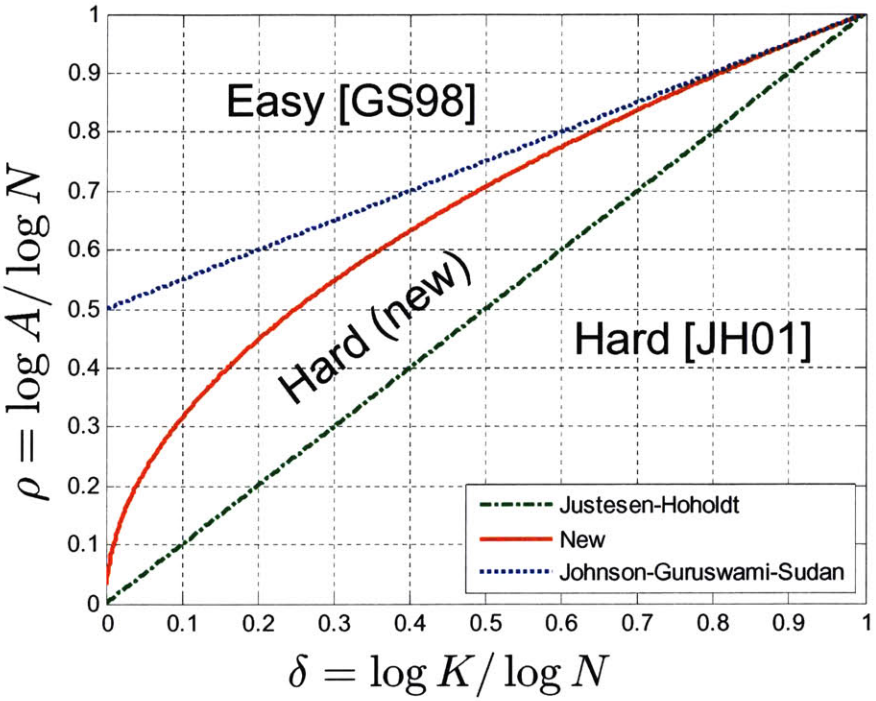

Figure 2-1: A comparison on double logarithmic scale, at $N \to \infty$, of the new bound (solid red line) with the Johnson-Guruswami-Sudan upper bound (dotted blue) and the previous Justesen-Høholdt lower bound (dashed green). The region where efficient list decoding algorithms exist is marked *Easy* and where efficient list decoding is provably infeasible is marked *Hard*. The status of the unmarked region (between the solid and dotted lines) is still to be resolved.

Chapter 3

Proof of the main result

3.1 The construction

We wish to show that there is a small ball that contains many codewords. We do this in two steps. First (Proposition 3.1.4), we relate the existence of a small ball with many codewords to the existence of a large family of polynomials with coefficients satisfying certain properties. Then (Lemma 3.1.5), by defining a suitable family of polynomials and using a counting argument, we show that the required large family of polynomials does exist.

As explained in the introduction, the key idea of our construction is the use of sparse polynomials that have many roots. Specifically, we work with polynomials whose roots form subspaces in a field. In this section, we define these polynomials formally and state the property that we need in our construction.

We view $\mathbb{K} = \mathbb{F}_{q^m}$ as a vector space of dimension m over the base field $\mathbb{F}_q$. In the proof of Corollary 2.1.1, we take q to be 2; however, in order to obtain good explicit constructions in Corollary 2.3.1, we need to let q grow with N.

Definition 3.1.1 (Subspace polynomials). *Let L be a subspace of $\mathbb{K}$ viewed as a vector space over $\mathbb{F}_q$. The subspace polynomial for L is defined by*

$$P_L(X) = \prod_{\ell \in L} (X - \ell).$$

We think of P_L as a function from $\mathbb{K}$ to $\mathbb{K}$, and for a set $S \subseteq \mathbb{K}$, use $P_L(S)$ to denote the image of S under P_L, that is, $P_L(S) = \{P_L(s) : s \in S\}$. The following proposition appears as [Ber68, Theorem 11.31] (see also [BSGH+04, Claim 8.15]). We include a proof for the sake of completeness.

Proposition 3.1.2. *Let L be a subspace of $\mathbb{K}$. Then, P_L has the form*

$$X^{q^{\dim L}} + \sum_{i=0}^{\dim L - 1} \alpha_i X^{q^i},$$

where $\alpha_i \in \mathbb{K}$.

Proof. Let $d = \dim L$, and consider the $d+1$ $\mathbb{F}_q$-linear functions from $\mathbb{K}$ to $\mathbb{K}$ defined by $f_i(x) = x^{q^i}$ for $i = 0, 1, \ldots, d$. The set of all $\mathbb{F}_q$-linear functions from L to $\mathbb{K}$ form a $\mathbb{K}$-vector space of dimension d. So, these $d+1$ linear functions must have a non-trivial dependency (when viewed as functions from L to $\mathbb{K}$), that is, there exist scalars $\alpha_0, \alpha_1 \ldots, \alpha_d \in \mathbb{K}$ (not all zero) such that $\sum_{i=0}^{d} \alpha_i f_i(x) = 0$ for all $x \in L$. Thus, there is a non-zero polynomial of the form $\sum_i \alpha_i X^{q^i}$ in $\mathbb{K}[X]$ of degree at most q^d that vanishes on L. We may assume that this polynomial is monic. But since L has q^d elements there is only one such polynomial (and its degree must be q^d), so it must be $P_L(X)$. □

Definition 3.1.3. *A family of polynomials $\mathcal{P} \subseteq \mathbb{K}[X]$ is said to be an (a, k)-family if*

1. *each polynomial in $\mathcal{P}$ has at least 'a' roots in $\mathbb{K}$, and*
2. *there is a polynomial P^* such that for all $P \in \mathcal{P}$, $P^* - P$ has degree at most k. We refer to the P^* as a* pivot *for the family.*

In the next proposition, we implement the first step of our plan: relate the existence of a ball with many codewords to the existence of a large (a, k)-family.

Proposition 3.1.4. *Let a, k and r be positive integers. Then, the following are equivalent.*

(a) There is a word $w : \mathbb{K} \to \mathbb{K}$ and r polynomials $P_1, P_2, \ldots, P_r$ of degree at most k such that for $i = 1, 2, \ldots, \ell$, $\mathsf{agree}(w, P_i) \geq a$.

(b) *There is an (a, k)-family of polynomials of size r whose pivot is the unique polynomial P_w of degree at most q^{m-1} that agrees with the word w on all points in $\mathbb{K}$.*

Proof. First, we show that (a) implies (b). Assume (a) holds. Observe that P_i agrees with w at the point x if and only if x is a root of the polynomial $P_w - P_i$. The required (a, k) family is then $\{P_w - P_i\}_{i=1}^{r}$.

Next, we show that (b) implies (a). Assume (b) holds, and fix an (a, k)-family $\{Q_1, Q_2, \ldots, Q_r\}$ with pivot Q^*. Define $w : \mathbb{F}_N \to \mathbb{F}_N$ by $w(x) = Q^*(x)$. Now, for $i = 1, 2, \ldots, r$, let $P_i = Q^* - Q_i$. Note that $w(x)$ and P_i agree on x if and only if x is root of $Q^* - P_i = Q_i$. Since Q_i has at least a roots, w and P_i agree on at least a points. Also, each P_i is a polynomial of degree at most k. Thus (a) holds. □

In the light of the above, in order to establish Theorem 2.0.1, it is enough to exhibit a large (q^u, q^v)-family of polynomials whenever $u \leq v$. We are now ready for the second step of our plan: exhibit such a family of large size. We need polynomials with many roots, say at least q^v. Also we need these polynomials to agree on all their coefficients for monomials of degrees more than q^u. We will concentrate on subspace polynomials corresponding to subspaces of dimension approximately v. These polynomials have degree N^v and the required N^v roots. Furthermore, by Proposition 3.1.2, they are sparse: they have only $v - u$ non-zero coefficients corresponding to monomials of degree more than q^u. By the pigeonhole principle a fraction at least $N^{-(v-u)}$ of these polynomials have the same top coefficients. Our main theorem will then follow by just counting the number of subspaces of dimension v. We now present this argument more formally.

Lemma 3.1.5 (Main lemma). *Let u and v be integers such that $0 \leq u \leq v \leq m$. Then, there is a (q^u, q^v)-family $\mathcal{P} \subseteq \mathbb{K}[X]$ of size at least $q^{(u+1)m-v^2}$. Furthermore, each polynomial in $\mathcal{P}$ is a subspace polynomial of degree q^v, and the pivot of the family has the form*

$$P^*(X) = X^{q^v} + \sum_{i=u+1}^{v-1} \alpha_i X^{q^i}.$$

Proof. Consider the subspace polynomials $P_L(X)$ where L is a v-dimensional subspace of $\mathbb{K}$. By Proposition 3.1.2, these polynomials have the form

$$X^{q^v} + \sum_{i=0}^{v-1} \alpha_i X^{q^i},$$

where $\alpha_i \in \mathbb{K}$. The number of subspaces of dimension v is

$$\frac{(q^m - 1)(q^m - q) \cdots (q^m - q^{v-1})}{(q^v - 1)(q^v - q) \cdots (q^v - q^{v-1})} \geq q^{v(m-v)},$$

and by the pigeonhole principle, for at least a fraction $q^{-m(v-u-1)}$ of these subspaces, their subspace polynomials have the same α_i for $i = u+1, u+2, \ldots, v$. Let this set of at least $q^{v(m-v)} \times q^{-m(v-u-1)} = q^{(u+1)m-v^2}$ subspaces be $\mathcal{L}$. Then,

$$\mathcal{P} = \{P_L : L \in \mathcal{L}\}$$

is the required (q^u, q^v)-family. □

Proof of Theorem 2.0.1 We combine Proposition 3.1.4 and Lemma 3.1.5. □

3.2 List decoding linearized polynomials beyond the Guruswami-Sudan bound

In this section, we give a list-decoding algorithm that list-decodes *linearized polynomials* (defined below) beyond the Guruswami-Sudan radius. This immediately implies combinatorial upper bounds on the number of linearized polynomials having certain agreement with any given received word. Our arguments in the previous chapters proceeded by proving lower bounds on the number of linearized polynomials having certain agreement with a particular received word. Thus different methods, i.e., going beyond subspace polynomials, are needed to prove the tightness of the Johnson-Guruswami-Sudan bound for Reed-Solomon codes. This algorithm also gives another explicit family of codes (after Parvaresh-Vardy [PV05] and Guruswami-Rudra [GR05a]), albeit very low rate, that can be list decoded beyond the Johnson bound.

Definition 3.2.1. *Let $\mathbb{F}_q$ be a field. An $\mathbb{F}_q$-linearized polynomial over $\mathbb{F}_{q^m}$ is a polynomial of the form*

$$\sum_{i=0}^{m-1} a_i X^{q^i}$$

where $a_i \in \mathbb{F}_{q^m}$ for each i.

If $f : \mathbb{F}_{q^m} \to \mathbb{F}_{q^m}$ is an $\mathbb{F}_q$-linearized polynomial, then f is $\mathbb{F}_q$-linear as a map between $\mathbb{F}_q$-vector spaces. In fact, every $\mathbb{F}_q$-linear map from $\mathbb{F}_{q^m}$ to $\mathbb{F}_{q^m}$ is the evaluation of some $\mathbb{F}_q$-linearized polynomial. By Proposition 3.1.2, subspace polynomials are also linearized polynomials.

We will need the Guruswami Sudan polynomial reconstruction algorithm [GS99].

Theorem 3.2.2. *Let $\mathbb{F}$ be a finite field. There is an algorithm GS−poly−reconstruct(S, K) that, given a multiset $S \subset \mathbb{F} \times \mathbb{F}$ and an integer K, finds all polynomials $f(X)$ of degree at most K such that:*

$$\sum_{(x,y)\in S} 1_{y=f(x)} > \sqrt{|S|K}$$

and runs in time poly$(|S|, |\mathbb{F}|)$.

The key idea for list decoding linearized polynomials is to exploit their $\mathbb{F}_q$-linearity. Suppose we were given a received word $r : \mathbb{F}_{q^m} \to \mathbb{F}_{q^m}$. Let us first find all linearized polynomials f which have $f(1) = a$. We know that for any such linearized polynomial f, for all $x \in \mathbb{F}_{q^m}$, $f(x+1) = f(x) + f(1) = f(x) + a$. Motivated by this we form the following 2 sets of points:

$$S_0 = \{(x, r(x)) : x \in \mathbb{F}_{q^m}\},$$

$$S_1 = \{(x, r(x+1) - a) : x \in \mathbb{F}_{q^m}\}.$$

Now consider any linearized polynomial $f(X)$ with $f(1) = a$. Let

$$V = \{(x, f(x)) : x \in \mathbb{F}_{q^m}\}$$

be its graph. We have $|V \cap S_0| = \text{agree}(f, r)$. By $\mathbb{F}_q$-linearity, $|V \cap S_1| = \text{agree}(f, r)$. Thus, (here $S_0 \cup S_1$ denotes the *multiset*)

$$\sum_{(x,y)\in S_0 \cup S_1} 1_{y=f(x)} = 2\text{agree}(f, r).$$

Thus, feeding the multiset $S_0 \cup S_1$ to the Guruswami-Sudan algorithm, we can recover all linearized polynomials of degree at most K with $f(1) = a$ as long as $\sqrt{|S_0 \cup S_1|K} \leq 2\text{agree}(f, r)$, i.e., as long as $\text{agree}(f, r) \geq \sqrt{q^m K/2} = \sqrt{NK/2}$. Now simply repeat this for all possible values of a. This list decodes linearized polynomials from an agreement of $\sqrt{NK/2}$. In comparison, the Guruswami-Sudan algorithm by itself would have required an agreement of $\sqrt{NK}$ for list decoding.

This basic idea is generalized in the following algorithm.

ALGORITHM 1 $(r : \mathbb{F}_{q^m} \to \mathbb{F}_{q^m}, K, c)$

1. Pick $x_1, \ldots, x_c \in \mathbb{F}_{q^m}$ linearly independent over $\mathbb{F}_q$.
2. For each $y = (y_1, \ldots, y_c) \in \mathbb{F}_{q^m}^c$, do:
   (a) For each $a = (a_1, \ldots, a_c) \in \mathbb{F}_q^c$, form the set:
   $$S_a = \left\{ \left(x, r(x - \sum_i a_i x_i) + \sum_i a_i y_i \right) : x \in \mathbb{F}_{q^m} \right\}$$
   (b) Define S to be the multiset $\bigcup_{a \in \mathbb{F}_q^c} S_a$.
   (c) Run $GS - poly - reconstruct(S, K)$ to get list $\mathcal{L}_y$.
3. Output $\mathcal{L} = \bigcup_{y \in \mathbb{F}_{q^m}^c} \mathcal{L}_y$.

Theorem 3.2.3. *Let q be a prime power and let $N = q^m$ for some m. For any $r : \mathbb{F}_N \to \mathbb{F}_N$, and any K, ALGORITHM 1 $(r : \mathbb{F}_N \to \mathbb{F}_N, K, c)$ returns a list of at most N^{c+2} polynomials that contains all linearized polynomials f of degree at most K with $\text{agree}(f, r) > \sqrt{NK/q^c}$. Furthermore, ALGORITHM 1 runs in time $O((qN)^c)$, with N^c calls to the Guruswami-Sudan polynomial reconstruction algorithm.*

Proof. Let f be a linearized polynomial of degree at most K with $\text{agree}(f, r) > \sqrt{NK/q^c}$. Consider the iteration of step 2 of the algorithm with $y = (f(x_1), f(x_2), \ldots, f(x_c))$. Then, in step 2(a), for each a, we get

$$S_a = \left\{ \left(x, r(x - \sum_i a_i x_i) + \sum_i a_i f(x_i) \right) : x \in \mathbb{F}_{q^m} \right\}.$$

Thus,

$$\sum_{(x,y)\in S_a} \mathbf{1}_{y=f(x)} = \sum_{x\in\mathbb{F}_{q^m}} \mathbf{1}_{r(x-\sum_i a_i x_i)+\sum_i a_i f(x_i)=f(x)}$$
$$= \sum_{x\in\mathbb{F}_{q^m}} \mathbf{1}_{r(x-\sum_i a_i x_i)=f(x-\sum_i a_i x_i)} = \text{agree}(f,r).$$

This implies that for $S =$ the multiset $\bigcup_a S_a$,

$$\sum_{(x,y)\in S} \mathbf{1}_{y=f(x)} = \sum_a \sum_{(x,y)\in S_a} \mathbf{1}_{y=f(x)} = \sum_a \text{agree}(f,r) = q^c \text{agree}(f,r) > \sqrt{(q^c N)K}. \tag{3.1}$$

When S is fed to the Guruswami-Sudan algorithm in step 2(c), since $|S| = q^c N$ and f is of degree K, Theorem 3.2.2 and (3.1) imply that f will be in the list $\mathcal{L}_y$, and hence in the final output list $\mathcal{L}$, as desired. The bounds on the running time are clear. □

Discussion: Theorem 3.2.3 shows that ALGORITHM 1 is a polynomial time list decoding algorithm for linearized polynomials. As a corollary, an upper bound on the list size is also obtained. Let $\epsilon > 0$ and set $c = 2\epsilon \log_q(N/K)$ in the theorem. For any $r : \mathbb{F}_{q^m} \to \mathbb{F}_{q^m}$, the number of linearized polynomials that have agreement at least $N\left(\frac{K}{N}\right)^{1/2+\epsilon}$ with r is at most $N^{2\epsilon \log_q(N/K)+2}$. Comparison with the discussion after Corollary 2.3.1 shows that both Theorem 3.2.3 and Corollary 2.3.1 are essentially optimal when $K = \Theta(N)$. Thus, in a certain sense, the list decoding lower bounds of Chapter 1 are optimal with regard to subspace polynomials, and different techniques are needed to prove the tightness of the Johnson-Guruswami-Sudan bound for Reed-Solomon Codes.

The above algorithm also shows that the (very low rate) subcode of the Reed-Solomon code $RS[N,K]$ consisting of linearized polynomials of degree at most K can be list decoded beyond the Johnson bound. This gives another explicit example of such codes after Parvaresh-Vardy [PV05] and Guruswami-Rudra [GR05a].

3.3 Open problems

- The proof of Theorem 2.0.1 uses a counting argument to find a set of coefficients that appear in many subspace polynomials. One method for obtaining an

explicit version of Corollary 2.1.1 (that is interesting in its own right) would be by gaining a better understanding of the *coefficients* of subspace polynomials.

- Recently, better constructions of list decodable codes based on RS codes were presented by Parvaresh & Vardy [PV05] and by Guruswami & Rudra [GR05a]. Could our techniques be useful in understanding the limitations of these new codes? In particular, an essential parameter in [PV05, GR05a] is the "folding" parameter (denoted M in [PV05] and m in [GR05a]) which is the number of field elements that comprise a single alphabet symbol in their codes. When $m = 1$ one obtains standard RS codes, and as m increases so does the efficient list-decoding radius (and the minimal required agreement decreases). For large m [GR05a] obtain efficient list decoding (with polynomial size lists) arbitrarily close to the rate of the code. Thus, we ask whether our techniques could imply combinatorial lower bounds for folded RS codes with a small number $m > 1$ of field elements per alphabet symbol?

Bibliography

[Ber68] Elwyn R. Berlekamp. *Algebraic Coding Theory.* Mc Graw-Hill, revised 1984 edition, 1968.

[BSGH+04] Eli Ben-Sasson, Oded Goldreich, Prahladh Harsha, Madhu Sudan, and Salil Vadhan. Robust PCPs of proximity, shorter PCPs and applications to coding. In ACM, editor, *Proceedings of the 36th Annual ACM Symposium on the Theory of Computing: Chicago, Illinois, USA, June 13–15, 2004*, pages 1–10, pub-ACM:adr, 2004. pub-ACM.

[BSGH+05] Eli Ben-Sasson, Oded Goldreich, Prahladh Harsha, Madhu Sudan, and Salil P. Vadhan. Short PCPs verifiable in polylogarithmic time. In *IEEE Conference on Computational Complexity*, pages 120–134, 2005.

[BSS05] Eli Ben-Sasson and Madhu Sudan. Simple PCPs with poly-log rate and query complexity. In *STOC '05: Proceedings of the 37th annual ACM Symposium on Theory of Computing*, pages 266–275, New York, NY, USA, 2005. ACM Press.

[CW04] Qi Cheng and Daqing Wan. On the list and bounded distance decodibility of Reed-Solomon codes. In IEEE, editor, *Proceedings: 45th Annual IEEE Symposium on Foundations of Computer Science: FOCS 2004, 17–19 October, 2004, Rome, Italy*, pages 335–341, 2004.

[DMS99] Ilya Dumer, Daniele Micciancio, and Madhu Sudan. Hardness of Approximating the Minimum Distance of a Linear Code. In *IEEE Symposium on Foundations of Computer Science*, pages 475–485, 1999.

[Eli57] Peter Elias. List decoding for noisy channels. Technical report, Research Laboratory of Electronics, MIT, 1957.

[GR05a] Venkatesan Guruswami and Atri Rudra. Explicit capacity-achieving list-decodable codes. *Electronic Colloquium on Computational Complexity (ECCC)*, (133), 2005.

[GR05b] Venkatesan Guruswami and Atri Rudra. Limits to list decoding Reed-Solomon codes. In *STOC '05: Proceedings of the thirty-seventh annual ACM symposium on Theory of computing*, pages 602–609, New York, NY, USA, 2005. ACM Press.

[GS99] Venkatesan Guruswami and Madhu Sudan. Improved Decoding of Reed-Solomon and Algebraic-Geometric Codes. In *IEEE Transactions on Information Theory*, volume 45, pages 1757–1767, 1999.

[GS01] V. Guruswami and M. Sudan. Extensions to the Johnson bound. 2001.

[GS02] Venkatesan Guruswami and Madhu Sudan. Reflections on Improved Decoding of Reed-Solomon and Algebraic-Geometric Codes, July 26 2002.

[Gur06] Venkatesan Guruswami. List Decoding in Average-Case Complexity and Pseudorandomness. Invited mini-survey to appear in ITW 2006, 2006.

[JH01] Jørn Justesen and Tom Høholdt. Bounds on list decoding of MDS codes. *IEEE Trans. Inform. Theory*, 47(4):1604–1609, 2001.

[Joh62] S. M. Johnson. A new upper bound for error-correcting codes. *IEEE Trans. on Information Theory*, 8:203–207, 1962.

[Joh63] S. M. Johnson. Improved asymptotic bounds for error-correcting codes. *IEEE Trans. on Information Theory*, 9:198–205, 1963.

[KY02] Aggelos Kiayias and Moti Yung. Cryptographic Hardness Based on the Decoding of Reed-Solomon Codes. In Peter Widmayer, Francisco Triguero Ruiz, Rafael Morales, Matthew Hennessy, Stephan Ei-

denbenz, and Ricardo Conejo, editors, *International Colloquium in Automata, Languages and Programming - ICALP 2002*, volume 2380 of *Lecture Notes in Computer Science*, pages 232–243. Springer, 2002.

[LN97] Rudolf Lidl and Harald Niederreiter. *Finite fields*, volume 20 of *Encyclopedia of Mathematics and its Applications*. Cambridge University Press, Cambridge, second edition, 1997. With a foreword by P. M. Cohn.

[Ore33] O. Ore. On a special class of polynomials. *Trans. Amer. Math. Soc.*, 35(3):559–584, 1933.

[Ore34] O. Ore. Contributions to the theory of finite fields. *Trans. Amer. Math. Soc.*, 36(2):243–274, 1934.

[PV05] Farzad Parvaresh and Alexander Vardy. Correcting errors beyond the guruswami-sudan radius in polynomial time. In *FOCS*, pages 285–294, 2005.

[RS60] I. S. Reed and G. Solomon. Polynomial Codes over Certain Finite Fields. *Journal of Society for Industrial and Applied Mathematics*, 8:300–304, 1960.

[Sud97] Madhu Sudan. Decoding of Reed Solomon Codes beyond the Error-Correction Bound. *Journal of Complexity*, 13(1):180–193, 1997.

[Woz58] J. M. Wozencraft. List decoding. *Quarterly progress report, Research Laboratory of Electronics, MIT*, 48:90–95, 1958.

[Xin03] Chaoping Xing. Nonlinear codes from algebraic curves improving the Tsfasman-Vladut-Zink bound. *IEEE Transactions on Information Theory*, 49(7):1653–1657, 2003.